Cross-database evaluation of audio-based spoofing detection systems

Pavel Korshunov and Sébastien Marcel

Biometrics group, Idiap Research Institute,
Martigny, Switzerland
pavel.korshunov@idiap.ch, sebastien.marcel@idiap.ch

Abstract

Since automatic speaker verification (ASV) systems are highly vulnerable to spoofing attacks, it is important to develop mechanisms that can detect such attacks. To be practical, however, a spoofing attack detection approach should have (i) high accuracy, (ii) be well-generalized for practical attacks, and (iii) be simple and efficient. Several audio-based spoofing detection methods have been proposed recently but their evaluation is limited to less realistic databases containing homogeneous data. In this paper, we consider eight existing presentation attack detection (PAD) methods and evaluate their performance using two major publicly available speaker databases with spoofing attacks: AVspoof[1] and ASVspoof[2]. We first show that realistic presentation attacks (speech is replayed to PAD system) are significantly more challenging for the considered PAD methods compared to the so called 'logical access' attacks (speech is presented to PAD system directly). Then, via a cross-database evaluation, we demonstrate that the existing methods generalize poorly when different databases or different types of attacks are used for training and testing. The results question the efficiency and practicality of the existing PAD systems, as well as, call for creation of databases with larger variety of realistic speech presentation attacks.

Index Terms: speaker anti-spoofing, presentation attacks, cross-database testing, open source

1. Introduction

Recent years have shown an increase in both the accuracy of biometric systems and their practical use. The application of biometrics is becoming widespread with fingerprint sensors in smartphones, automatic face recognition in social networks and video-based applications, and speaker recognition in phone banking and other phone-based services. The popularization of the biometric systems, however, exposed their major flaw — high vulnerability to spoofing attacks [1]. A fingerprint sensor can be easily tricked with a simple glue-made mold, a face recognition system can be accessed using a printed photo, and a speaker recognition can be spoofed with a replay of pre-recorded voice. The ease with which a biometric system can be spoofed demonstrates the importance of developing efficient anti-spoofing systems that can detect both known (conceivable now) and unknown (possible in the future) spoofing attacks.

In this paper, we focus on the spoofing attack detection or presentation attack detection (PAD) systems in the context of voice biometrics. Although the research in this area is far from being matured, several approaches have been proposed recently, mostly focusing on feature extraction component of anti-spoofing systems. A survey by Wu *et al.* [2] provides a comprehensive overview of the spoofing attacks and the currently available anti-spoofing methods. These methods use features mostly based on the audio spectrogram, such as spectral- and cepstral-based features [3], phase-based features [4], the combination of amplitude and phase features [5], and audio quality based features [6]. Also, a higher computational layer can be added, for instance, Alegre *et al.* [7] proposed to use histograms of local binary patterns (LBP), which can be computed directly from a set of pre-selected spectral, phase-based, or other features. Most of these features are used successfully in speaker verification systems already, so, naturally, they are first to be proposed for anti-spoofing systems as well. However, to demonstrate that the anti-spoofing methods work, they need to be evaluated on a dataset with a set of comprehensive spoofing attacks that pose a real threat to the state of the art speaker verification systems [8].

Although, the number of speech databases with spoofing attacks is significantly limited, two major databases were created recently: ASVspoof[1] [9], as part of the 2015 Interspeech anti-spoofing challenge, and AVspoof[2] [10]. Most of the available work on speaker anti-spoofing, essentially, focuses on detecting synthetic speech, such as voice conversion, speech synthesis, and artificial signals [2]. The synthetically generated attacks are assumed to be fed into a verification system directly bypassing its microphone, hence, they are coined as 'logical access' attacks [9]. This type of attacks constitute ASVspoof database. The most practical *replay attacks*, which are formally defined as *presentation attacks* by ISO standardization committee [11], received considerably less attention, since, until now, there was no public dataset with such attacks. That is why AVspoof database [10] is of great interest, since it is the first database that contains several types of replay attacks.

To overcome the lack of evaluations of the currently available spoofing attack detection methods on presentation attacks and to find out how well these methods generalize across different types of attacks, we need an extensive cross-database evaluations of the state-of the art. By taking the recent work of Sahidullah *et al.* [12], which benchmarked several anti-spoofing systems, as a starting point, we have selected several well-performing methods and developed their open source implementation[3] based on a well-known Bob framework [13][4]. Hence, we have implemented: GMM-based classifier using cepstral-based features with rectangular (RFCC), mel-scaled triangular (MFCC) [14], inverted mel-scaled triangular (IMFCC), and linear triangular (LFCC) filters [15], spectral flux-based features (SSFC) [16], subband centroid frequency (SCFC) [17], and subband centroid magnitude (SCMC) [17] features, as well as, features of LBP histograms computed us-

[1] https://www.idiap.ch/dataset/avspoof
[2] http://datashare.is.ed.ac.uk/handle/10283/853
[3] Source code: https://pypi.python.org/pypi/bob.paper.interspeech_2016
[4] http://idiap.github.io/bob/

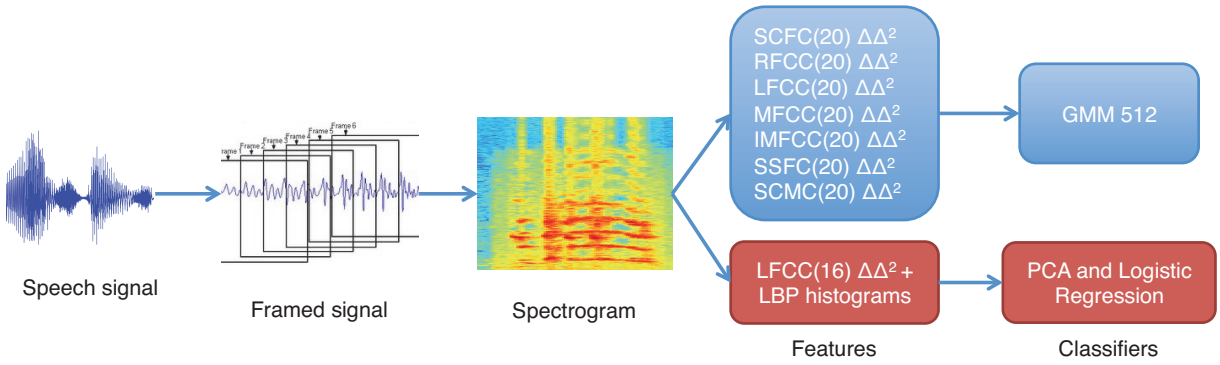


Figure 1: Processing flow of considered PAD systems.

ing LFCC [7] with a PCA-based dimensionality reduction and logistic regression classifier. Figure 1 illustrates the processing flow of the implemented PAD systems.

We demonstrate the effect of presentation (or replay) attacks on the selected anti-spoofing systems compared to synthetically generated attacks. And to understand how well these systems generalize across different databases and different types of data, we conducted an extensive cross-database evaluation by (i) training the systems on data from one database and testing them on data from another database, and (ii) training them on one type of data ('logical access' or presentation attacks) and testing them on another type of data.

Therefore, the main contributions of this paper include (i) open source implementation of eight state of the art speaker spoofing detection systems, (ii) evaluation of these systems on presentation attacks from AVspoof database, and (iii) comparative performance and cross-database testing using AVspoof and ASVspoof databases.

2. Spoofing detection methods

We have selected several state of the art methods for spoofing attacks detection in speech (please see Figure 1 for an illustration), which were recently evaluated by Sahidullah *et al.* [12] on ASVspoof database. Taking already benchmarked approaches allow us to compare and cross-validate our open source implementations[3] with the results presented in [12].

By following [12], we adopted GMM-based classifier (two models for real and attacks), since it led to a better performance compared to SVM. We selected four cepstral-based features with rectangular (RFCC), mel-scale triangular (MFCC) [14], inverted mel-scale triangular (IMFCC), and linear triangular (LFCC) filters [15]. These features are computed from a power spectrum (power of magnitude of 512-sized FFT) by applying one of the above filters of a given size (we use size 20 as per [12]). We also implemented spectral flux-based features (SSFC) [16], which are Euclidean distances between power spectrums (normalized by the maximum value) of two consecutive frames, subband centroid frequency (SCFC) [17], and subband centroid magnitude (SCMC) [17] features. A discrete cosine transform (DCT-II) is applied to all above features, except for SCFC, and first 20 coefficients are taken.

For the comparison with the above systems, we also implemented features using histograms of uniform local binary patterns (uLBP(8,1)) [7] computed from LFCCs (16 coefficients, deltas, and double-deltas [18]), which is one of the first methods proposed for anti-spoofing detection in speech. However, instead of using a histogram intersection for classification between real data and attacks, to increase the accuracy of detection closer to practical values, we used logistic regression with prior PCA-based dimensionality reduction.

Before computing selected features, a given audio sample is first split into overlapping 20ms-long speech frames with 10ms overlap. The frames are pre-emphasized with 0.97 coefficient and pre-processed by applying Hamming window. Then, for all features, except for LBP-based features [7], we compute deltas and double-deltas [18] and keep only these features (40 in total) for the classifier. We kept only deltas and double-deltas, because [12] reported that static features degraded performance of PAD systems.

3. Spoofing databases

For our cross-database benchmarking of selected anti-spoofing methods, we have selected two recent and the most comprehensive to date publicly available databases with speech spoofing attacks: ASVspoof and AVspoof.

3.1. ASVspoof

ASVspoof[1] database was created for a 2015 Interspeech anti-spoofing challenge [9]. It contains genuine speech data from 106 speakers (45 male and 61 female), while spoofed speech was generated using speech synthesis and voice conversion algorithms. In total, database has 10 spoofing attacks, five of which are considered 'unknown', since they appear in the evaluation set only and their nature was not known to the participants at the time of the evaluation. However, it is important to note that ASVspoof database was built with an assumption that the attackers have a direct access (or what the authors called 'logical access') to a verification system. It means that the database does not contain more realistic presentation attacks [11] and hence, basically, is a database of synthetic speech. For more details on ASVspoof database, please refer to [9].

3.2. AVspoof

To our knowledge, the largest publicly available database containing speech presentation attacks is AVspoof [10][2].

AVspoof database contains real (genuine) speech samples from 44 participants (31 males and 13 females) recorded over the period of two months in four sessions, each scheduled several days apart in different setups and environmental conditions

Table 1: Performance of PAD systems on ASVspoof [9] and AVspoof [10] with results in *[12] EER (%)* column copied from [12].

PAD system	**ASVspoof (Eval)**				**AVspoof (Eval)**	
	[12] EER (%)		**Bob[3] EER (%)**		**Bob[3] HTER (%)**	**Bob[3] HTER (%)**
	Known	Unknown	Known	Unknown	LA	PA
SCFC(20) $\Delta\Delta^2$, GMM512	0.07	8.84	0.10	5.17	0.00	5.15
RFCC(20) $\Delta\Delta^2$, GMM512	0.12	1.92	0.12	1.32	0.03	2.70
LFCC(20) $\Delta\Delta^2$, GMM512	0.11	1.67	0.13	1.20	0.00	5.00
MFCC(20) $\Delta\Delta^2$, GMM512	0.39	3.84	0.46	2.93	0.00	5.34
IMFCC(20) $\Delta\Delta^2$, GMM512	0.15	1.86	0.20	1.57	0.01	3.76
SSFC(20) $\Delta\Delta^2$, GMM512	0.30	1.96	0.23	1.60	0.70	4.17
SCMC(20) $\Delta\Delta^2$, GMM512	0.17	1.71	0.18	1.37	0.01	3.24
uLBP(8,1), PCA and LR	N/A	N/A	9.95	21.30	0.65	5.89

Table 2: Cross database testing on ASVspoof [9] and AVspoof [10] databases of PAD systems (open source implementations[3]).

PAD system	**HTER (%)**			
	ASVspoof (Train/Dev)		**AVspoof-LA (Train/Dev)**	
	AVspoof-LA (Eval)	**AVspoof-PA (Eval)**	**ASVspoof (Eval)**	**AVspoof-PA (Eval)**
SCFC(20) $\Delta\Delta^2$, GMM512	1.43	6.48	19.99	7.56
RFCC(20) $\Delta\Delta^2$, GMM512	34.93	38.54	25.58	13.20
LFCC(20) $\Delta\Delta^2$, GMM512	0.71	10.58	18.44	8.40
MFCC(20) $\Delta\Delta^2$, GMM512	1.87	9.82	10.13	5.15
IMFCC(20) $\Delta\Delta^2$, GMM512	2.28	46.49	21.80	49.57
SSFC(20) $\Delta\Delta^2$, GMM512	34.64	41.68	43.50	36.26
SCMC(20) $\Delta\Delta^2$, GMM512	1.23	12.16	22.99	7.97
uLBP(8,1), PCA and LR	58.38	53.00	44.34	49.10

such as background noises. The first session was recorded in the most controlled conditions. Speech samples were recorded using three devices: laptop using microphone AT2020USB+, Samsung Galaxy S4 phone, and iPhone 3GS.

From the recorded genuine data, two major types of attacks were created for AVspoof database: 'logical access' attacks, similar to those in ASVspoof database [9], and presentation attacks, as they are defined by ISO standardization committee [11]. 'Logical access' attacks are generated using (i) a statistical parametric-based speech synthesis algorithm [19] and (ii) a voice conversion algorithm from Festvox[5].

When generating presentation attacks, the assumption is that a verification system is installed on a laptop (with an internal built-in microphone) and an attacker is trying to gain access to this system by playing back to it a pre-recorded genuine data or an automatically generated synthetic data using some playback device. In AVspoof database, presentation attacks consist of (i) direct replay attacks when a genuine data is played back using a laptop with internal speakers, a laptop with external high quality speakers, Samsung Galaxy S4 phone, and iPhone 3G, (ii) synthesized speech replayed with a laptop, and (iii) converted voice attacks replayed with a laptop. For more details on AVspoof database, please refer to [10].

4. Evaluation results

Open source implementations[3] of the selected spoofing detection methods (see Section 2 for details) are evaluated by computing a half total error rate (HTER) on the evaluation set of a database, which is a mean of false acceptance rate (FAR) and false reject rate (FRR) computed using a threshold found from the development set (see [8] for more details).

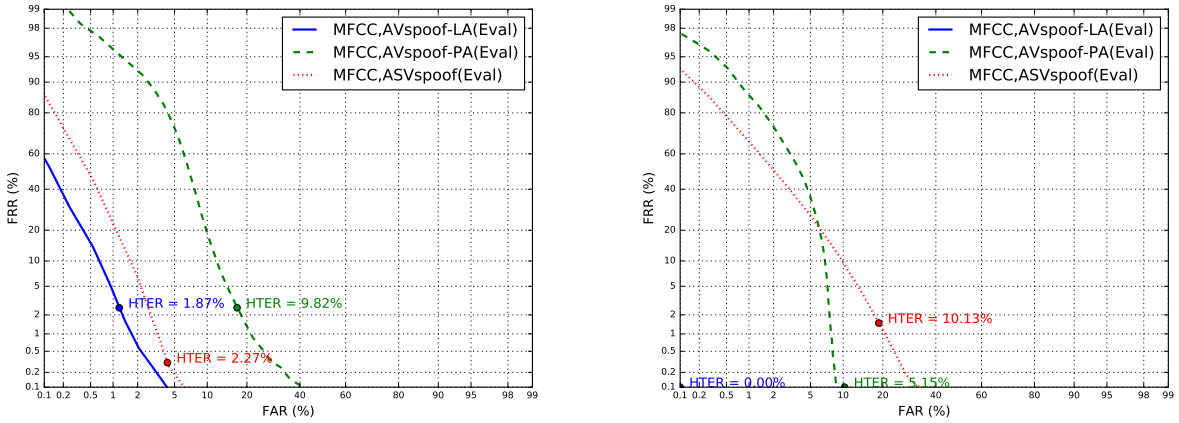
[5] http://festvox.org/

4.1. Comparative evaluation ASVspoof vs. AVspoof

The results of evaluating the selected methods on different types of attacks from ASVspoof and AVspoof databases are presented in Table 1. These results are obtained by training each anti-spoofing method (i.e., a PAD system) on a training (*Train*) set of a given database, then tuned on development (*Dev*) set and tested on evaluation (*Eval*) set of the same database.

In Table 1, the results for 'known' and 'unknown' attacks (see columns titled as *Bob EER (%)* to indicate our Bob-based open source implementation) of *Eval* set of ASVspoof are presented separately. Thus, our results can be compared with results reported in Table 4 of [12] (Matlab implementation), which we also included in Table 1 for convenience (columns titled as *[12] EER (%)*). Also, equal error rates (EER) are shown for these sets to keep the results comparable with [12]. We can observe that both implementations lead to similar results for 'known' attacks, while our Bob-based system shows smaller error rates for 'unknown' attacks.

For AVspoof database, only the results for our Bob-based implementation are shown in Table 1, since AVspoof database was not used in [12]. Since AVspoof contains both 'logical access' (LA for short) and presentation attacks (PA), the results for these two types of attacks are presented separately, to allow comparing the performance on ASVspoof database (it has 'logical access' attacks only) with an AVspoof-LA attacks. From these results, we can note that (i) LA set of AVspoof is less challenging compared to ASVspoof for all methods, except for SSFC-based, and (ii) presentation attacks are significantly more challenging compared to LA attacks. The increase in HTER for almost all PAD systems in more than 100 times for PA compared to LA attacks of AVspoof means that presentation attacks, besides emulating a more realistic spoofing scenario, pose a serious threat to the state of the art systems and need to be considered in all future evaluations of anti-spoofing systems.



(a) *Train* and *Dev* sets of ASVspoof

(b) *Train* and *Dev* sets of AVspoof-LA

Figure 2: DET curves for *MFCC(20)* $\Delta\Delta^2$, *GMM512* PAD system when it is trained on one database and evaluated on another.

Table 3: Per attack results for RFFC and MFCC-based systems.

	Attacks	**ASVspoof (Train)**		**AVspoof-LA (Train)**	
		MFCC	**RFCC**	**MFCC**	**RFCC**
AVSpoof-LA	SS	1.45	34.78	0.00	1.07
(Eval)	VC	1.88	34.93	0.00	0.00
AVSpoof-PA	SS	3.51	35.14	24.91	40.09
(Eval)	SS-HQ	22.93	39.11	11.52	36.79
	VC	6.21	35.65	0.90	9.40
	VC-HQ	6.30	35.40	1.12	10.32
	RE	48.25	72.20	49.98	49.99

4.2. Cross-database evaluation

Table 2 presents the cross-database results when a given PAD system is trained and tuned using training and development sets from one database but is tested using evaluation set from another database. For instance, results in the second column of the table are obtained by using training and development sets from ASVspoof database but evaluation set from AVspoof-LA. Also, we evaluated the effect of using one type of attacks (e.g., 'logical access' from AVspoof-LA) for training and another type (e.g., presentation attacks of AVspoof-PA) for testing (the results are in the last column of the table).

From the results in Table 2, we can note that all methods generalize poorly across different datasets with HTER reaching 50%. It is also interesting to note that even similar methods, for instance, RFCC and MFCC-based, have very different accuracy in cross-database testing, even though they showed less drastic difference in single-database evaluations (see Table 1). To understand the cause for such drastic difference in performances, we split the results for RFCC and MFCC-based methods for each attack of AVspoof database separately, as shown in Table 3. The rows of the table indicate the attacks, including 'logical access' speech synthesis (*SS*) and voice conversion (*VC*) in AVspoof-LA, and replay or presentation attacks of speech synthesis (*SS*), speech synthesis replayed with high quality speakers (*SS-HQ*), voice conversion (*VC*), voice conversion replayed with high quality speakers (*VC-HQ*), and replay attacks with laptop and phones (*RE*). We can note that the main difference between RFCC and MFCC-based methods is in SS and VC attacks, while both methods perform almost equally poorly on RE attacks. It means MFCC-based system is more sensitive to synthetic attacks and, since such synthetic attacks constitute the majority of the overall attacks, it led to the overall low HTER for MFCC-based system. But none of the systems work well on *RE* attacks.

To provide more details on how MFCC-based PAD system performs on different databases, we plot its detection error tradeoff (DET) curves in Figure 2. Figure 2a shows DET curves when the method is trained on ASVspoof database and Figure 2b shows when it is trained on AVspoof-LA. In both plots, we can notice that the PAD system is misbalanced: a small decrease in false acceptance rate of spoofing attacks has a cost of having large increase in false rejection rate. This is especially true when the system is evaluated using attacks from AVspoof-PA, which again emphasizes the challenge that presentation attacks pose to the current PAD systems.

5. Conclusion

In this paper, we conducted a cross-database evaluation of several state of the art speech spoofing attacks detection methods implemented as open source. We used two recent comprehensive databases with speech spoofing attacks: ASVspoof ('logical access' attacks only) and AVspoof ('logical access' and presentation attacks). The evaluation results demonstrated that all methods generalize poorly, especially, when they are trained on 'logical access' attacks and tested on more realistic presentation attacks, which means a new and more practically applicable anti-spoofing methods need to be developed. Databases with larger variety of attacks in both training and testing sets are also needed.

6. Acknowledgements

This work was partially funded by Google Inc., EU BEAT project (#284989), and Norwegian SWAN project. The authors would also like to thank Sahid Sahidullah for providing implementation details of the features from [12] paper.

7. References

[1] S. Marcel, M. S. Nixon, and S. Z. Li, *Handbook of Biometric Anti-Spoofing: Trusted Biometrics Under Spoofing Attacks*. Springer-Verlag London, 2014.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, Feb. 2015.

[3] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," in *INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 239–243.

[4] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2280–2290, Oct. 2012.

[5] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2062–2066.

[6] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2077–2081.

[7] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Arlington, VA, Sep. 2013, pp. 1–8.

[8] I. Chingovska, A. Anjos, and S. Marcel, "Biometrics evaluation under spoofing attacks," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2264–2276, Dec. 2014.

[9] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," pp. 2037–2041, Sep. 2015.

[10] S. Kucur Ergunay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, Sep. 2015.

[11] ISO/IEC JTC 1/SC 37 Biometrics, "DIS 30107-1, information technology – biometrics presentation attack detection," American National Standards Institute, Jan. 2016.

[12] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2087–2091.

[13] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *ACM international conference on Multimedia (ACMMM)*, Oct. 2012.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[15] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[16] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Apr. 1997, pp. 1331–1334.

[17] P. N. Le, E. Ambikairajah, J. Epps, V. Sethu, and E. Choi, "Investigation of spectral centroid features for cognitive load classification," *Speech Commun.*, vol. 53, no. 4, pp. 540–551, Apr. 2011.

[18] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 6, pp. 871–879, Jun. 1988.

[19] H. Zen, K. Tokuda, and A. W. Black, "Review: Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.